DE-FG01-96ER62327.A000

DOE/ER/62327--T1

DATABASE MANAGEMENT RESEARCH FOR THE
HUMAN GENOME PROJECT

Progress Report
for period: 07/01/96 - -3/15/97

Nathan Goodman, Ph.D., Principal Investigator

The Jackson Laboratory
600 Main Street
Bar Harbor, ME 04609
(207) 288-6158

RECEIVED
MAR 17 1997
OSTI

March 1997

Prepared for
THE U.S. DEPARTMENT OF ENERGY
AWARD NO. DE-FG02-96ER62327.A000

MASTER

NOTICE:

This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Department of Energy, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness or any information, apparatus, product or process disclosed or represents that its use would not infringe privately-owned rights.

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

DISCLAIMER

**Portions of this document may be illegible
in electronic image products. Images are
produced from the best available original
document.**

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, make any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

This document reports on progress for the first 8 months of my research project entitled "Database Research for the Human Genome Project". This project continues the work I began at Whitehead Institute on a previous DOE grant entitled "A Freely Sharable Database Management System Designed for Use in Component-Based, Modular, Genome-Informatics Systems".

The goal of this research is to develop software and techniques that make it easier to create databases for a variety of purposes within the Human Genome Project. Our main technical approach has been to develop software that works in conjunction with database management systems (DBMSs), and which augments the facilities of the DBMSs in ways that are useful for genomics. At Whitehead, my colleagues Drs. Lincoln Stein and Steve Rozen, and I used this approach to develop LabBase, software that augments the ObjectStore DBMS to make it easier to create databases for high throughput genome laboratories.

Since moving to the Jackson Laboratory at the start of the current reporting period, I have continued my collaboration with Drs. Stein and Rozen on LabBase, and we have jointly produced a major new release of this software. This new release of LabBase has two major advantages over the previous version.

First, the new release runs on the Sybase relational DBMS, rather than ObjectStore. This is important, because Sybase is the most widely used DBMS product in the Human Genome Project. By moving LabBase from ObjectStore to Sybase, we will make it much easier for other investigators to use our software. In addition, we believe the performance of this version will scale more readily to the levels needed for mega-sequencing projects. We are in the process of porting this software to another relational DBMS, namely. This is important, because Oracle is the DBMS-of-choice for at least two major sequencing centers and is almost universally used in the pharmaceutical industry.

Second, the new release offers more complete data modeling features than the previous version and is suitable for more kinds of genomic databases. The data modeling features we added were strongly inspired by the eminently successful ACEDB data model. As in ACEDB, our data model provides direct support for inter-object links, multi-valued attributes, and nested structures (sub-objects). Also as in ACEDB, our model provides straightforward ways to store common genomic data types such as long sequences.

The new release is functionally complete and has entered quality-assurance. Work on performance enhancement is also underway. A draft reference manual has been written, and the final version is almost complete. Once the software passes through quality-assurance and the manual is complete, we will install the software at 3-5 beta sites.

The software is written in Perl and consists of about 4000 lines of code. The software has three major modules. One module contains all logic that is specific to a particular DBMS. The second module contains DBMS-independent logic that implements all LabBase functions operating on "user data", e.g. functions to store and retrieve database objects; this software, of course, utilizes the services of the DBMS-specific module. The third module contains DBMS-independent logic that operates on "system data", primarily the database schema; this module utilizes the services of the second module to access schema information in the database.

The next stage of the project will be to port the software to Informix Universal Server, an object-relational DBMS that supersedes Illustra. This is a product that has been adopted by at least two genome laboratories: Gerry Rubin's Berkeley Drosophila Genome Center, and Mike Cherry's Arabidopsis Genome Center at Stanford. I expect this technology to offer several major advantages for LabBase. First, it will allow us to simplify the LabBase software which is always a good thing from an engineering perspective. Second, it will provide a high level, SQL-based query language, which will make it easier for programmers to create software that uses the database, and for end-users to access the database via *ad hoc* queries. Third, it will give us a better model for "extensibility", by which I mean the ability for programmers to add new features to the data model provided by the basic LabBase software; experience with ACEDB, for example at the Berkeley Drosophila Genome Center, suggests that extensibility is a key feature for software of this sort to be usable across a wide range of genome applications.

In the Final Report for the Whitehead project, I discussed our efforts to implement a version of LabBase on the freely available *libdb* package from Berkeley, and I explained how this effort faltered because necessary enhancements to *libdb* could not be developed in time. Subsequently, and through no effort on our part, the developers of *libdb* were able to raise commercial funds to complete the needed enhancements. The *libdb* developers are still interested in working with us to create a version of LabBase on their package. This offers us the opportunity to resume our LabBase/*libdb* effort and thereby obtain a robust, freely available version of the complete system. We have not yet decided whether to pursue this opportunity.